A Preliminary Analysis of the InfiniPath and XD1 Network Interfaces

Ron Brightwell **Doug Doerfler** **Keith Underwood**

Sandia National Laboratories

Center for Computation, Computers, Information, and Mathematics

Workshop on Communication Architectures for Clusters

April 25, 2006

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

Sandia National Laboratories

Exploring Recent Trend in Network Interfaces

- Leveraging commodity technology while providing some hardware innovation to deliver increased performance
  - Both PathScale InfiniPath and Cray's Rapid Array Interconnect (RAI) leverage InfiniBand transport layer and HyperTransport
- InfiniPath introduces more radical change
  - Move processing from the network interface to the host CPU(s)

Sandia National Laboratories

Extends Previous Evaluation

- Cluster 2004 paper provided analysis of Elan-4 and IB
- This paper does the same type of analysis for InfiniPath and RAI
- We look at five areas
  - Capabilities
  - Programming interface
  - Connection establishment
  - Memory registration
  - Progress, offload, and overlap

Sandia National Laboratories

PathScale InfiniPath

- Few technical details describing implementation
- Process has been to guess and let Greg Lindahl correct
- Main philosophy is to move functions typically performed by a relatively slow NIC processor to a much faster host processor
- No transmit DMA engines on the interface
  - Host processor must move data from host memory to NIC memory
- NIC recognizes incoming write and streams data onto the network
- Receive-side writes incoming messages into host memory and records where they have been written
- Destination is either explicit or anonymous
- Host is responsible for recognizing errors and performance reliability and flow control functions

Sandia National Laboratories

Cray's Rapid Array Interconnect (RAI)

- Even fewer published technical details
- RAI has processors on the NIC to offload and accelerate core network functions to unburden the host and provide overlap
- Unknown how much these units differ from traditional IB NICs
- Small MPI messages done with memory-to-memory copies
- Transmit DMA engine for large transfers

Sandia National Laboratories

Programming API

- InfiniPath
  - Write directly into mapped NIC memory
  - Supports OpenIB API
- RAI
  - Similar to VIA-based APIs like VAPI and uDAPL
- Neither support the ability to do MPI tag matching

Sandia National Laboratories

Connections

- InfiniPath
  - Connectionless
  - No concept of a queue pair
- RAI
  - Explicit connection establishment
  - VIA/IB queue pairs
  - Application memory must be committed to each endpoint

Sandia National Laboratories

Memory Registration

- InfiniPath
  - No registration required for transmits
  - Zero-copy receives require explicit memory registration
- RAI
  - Explicit registration for send, receive, and RDMA buffers

Sandia National Laboratories

Progress, Offload, and Overlap

- Progress
  - MPI posted receive queue in user space means neither InfiniPath nor RAI have independent progress for long message transfers
- Offload
  - Neither NIC does offload
    - InfiniPath approach directly conflicts
- Overlap
  - RAI supports overlap for RDMA, but is hampered for MPI
  - InfiniPath approach directly conflicts

Sandia National Laboratories

Test Platforms

	Emerald	Red Squall	Thunderbird	Cray XD1
Interconnect	4x InfiniPath	Elan-4	4x InfiniBand	Dual 4x IB
Host Interface	HyperTransport	PCI-X	x8 PCI-E	HyperTranspot
Peak Link BW	2 GB/s	2.133 GB/s	2 GB/s	4 GB/s
Host Interfce BW	6.4 GB/s	1.064 GB/s	4 GB/s	3.2 GB/s
Host CPU(s)	4 2.2 GHz Opteron	2 2.2 GHz Opteron	2 3.4 GHz EM64T	2 2.2 GHz Opteron
Memory Speed	Dual DDR-400	Dual DDR-333	Dual DDR-400	Dual DDR-400
OS	RHEL-4	SUSE-9.1 Pro	SUSE-9.1 Pro	SLES 9
Compilers	PathScale 2.2	PathScale 2.1	PathScale 2.1	PGI 6.0.5
MPI	InfiniPath 1.1	QsNet 1.24-43	MVAPICH 0.92	MPICH 1.2.6
Nodes	144	256	4096	72

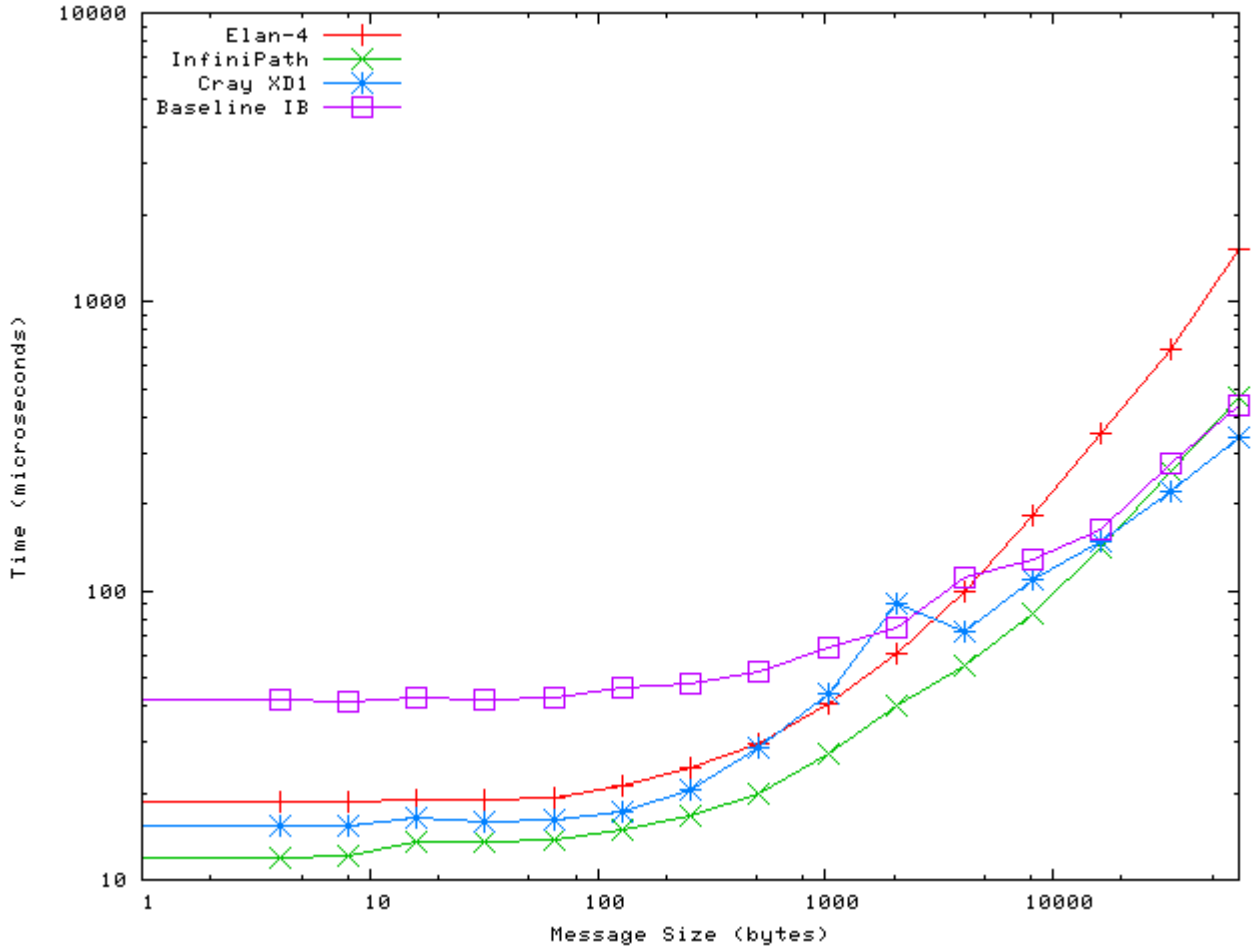
Sandia National Laboratories

Micro-Benchmarks and Application Tests

- Micro-benchmarks
  - Pallas MPI benchmark suite
  - OSU streaming bandwidth
    - 160 outstanding sends
    - Also used to calculate message rate
  - COMB benchmark suite
    - Polling method
    - Used to calculate CPU availability
- Application
  - LAMMPS molecular dynamics simulation
    - 2001 Fortran version
    - 2005 C++ version

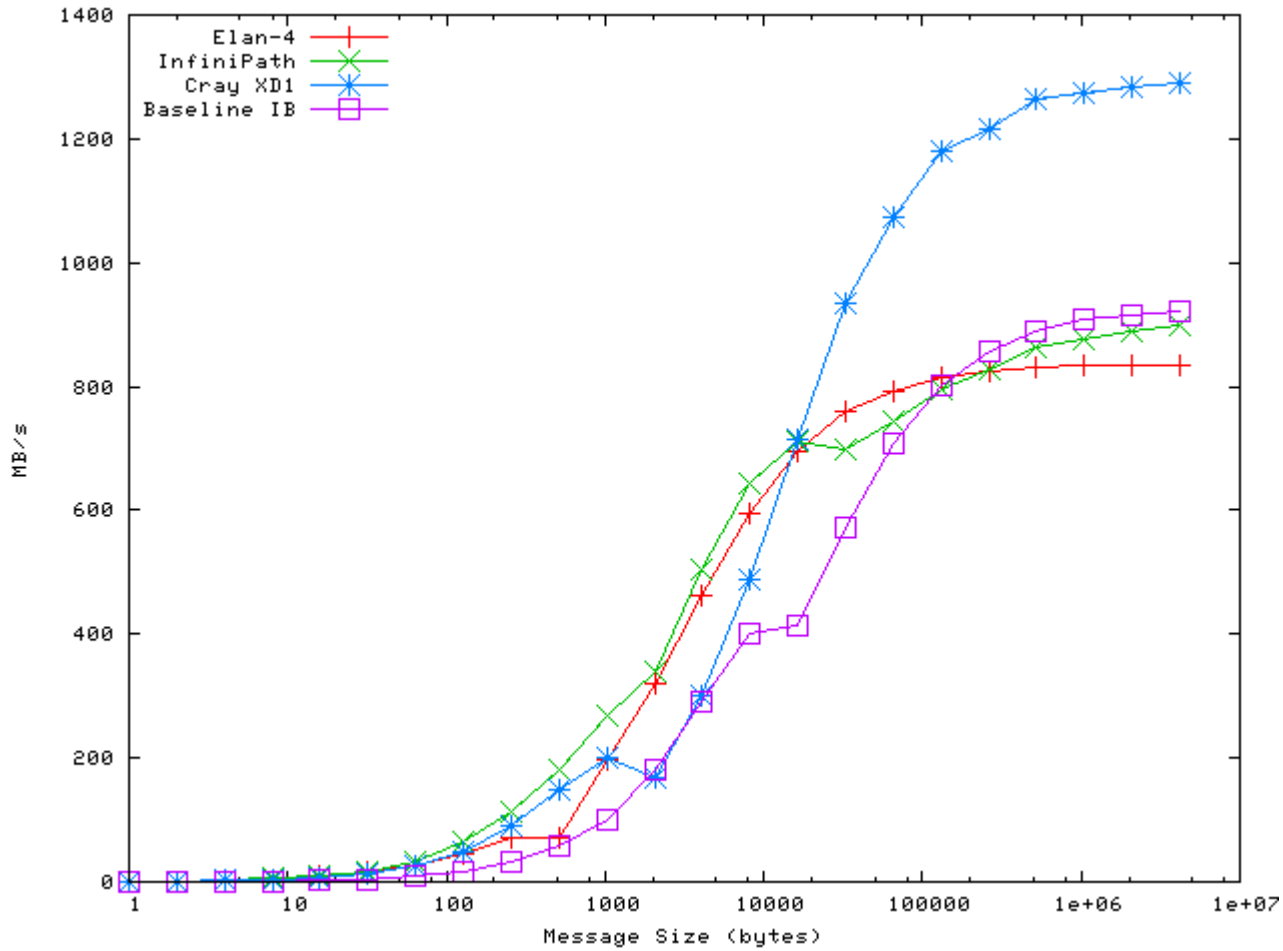
Sandia National Laboratories

Ping-Pong Latency



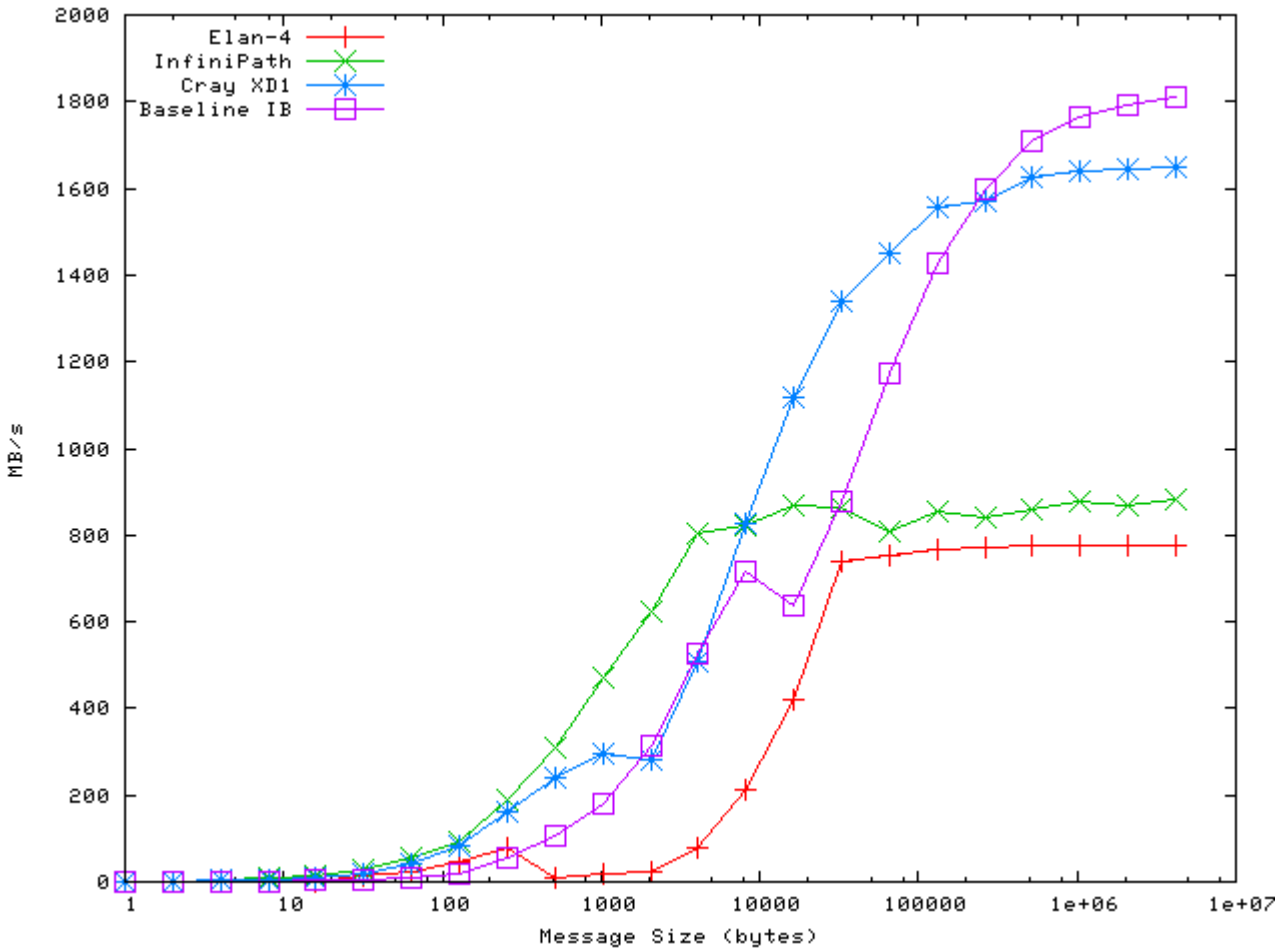
Sandia National Laboratories

Ping-Pong Bandwidth



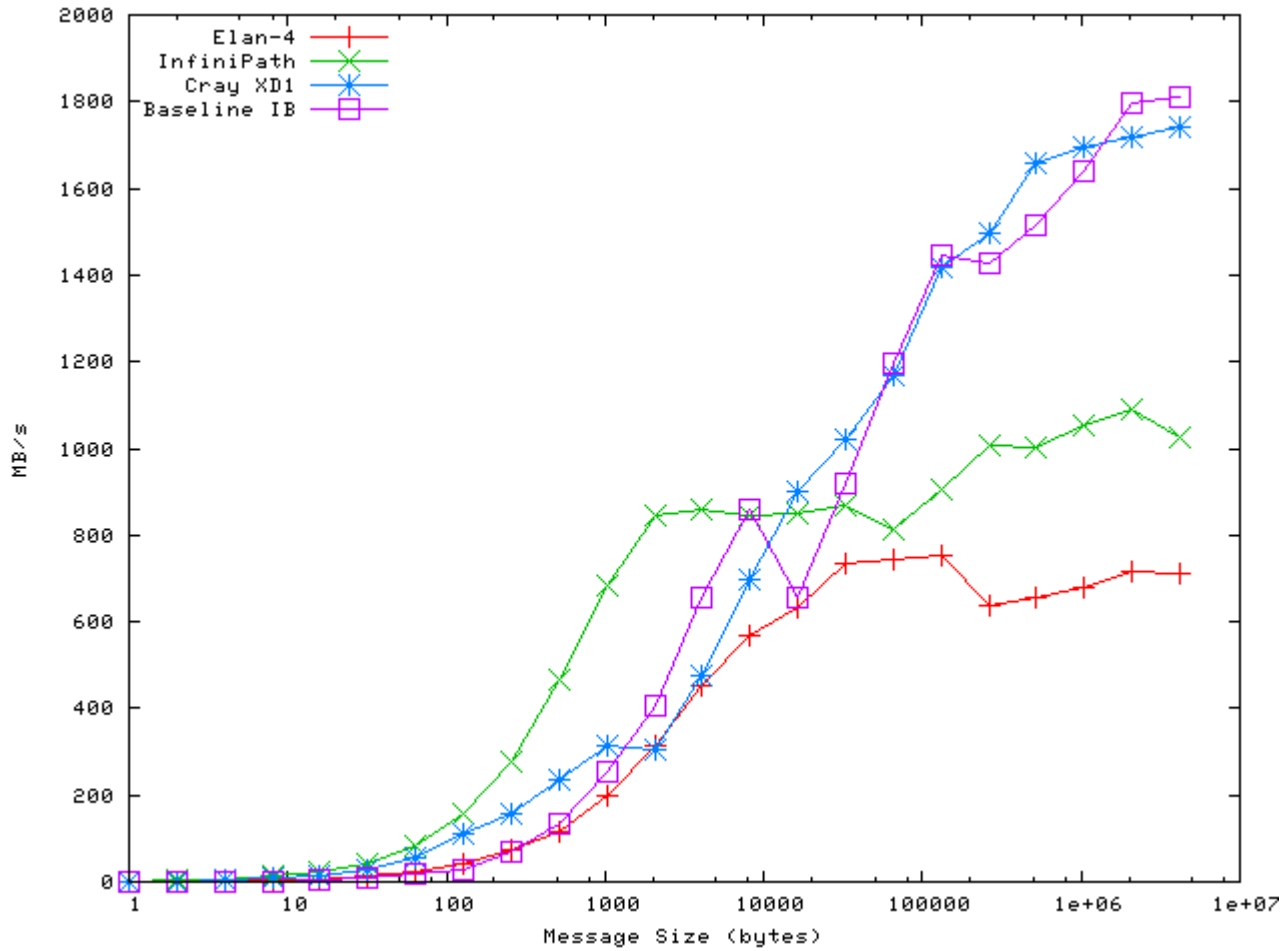
Sandia National Laboratories

Send-Receive Bandwidth



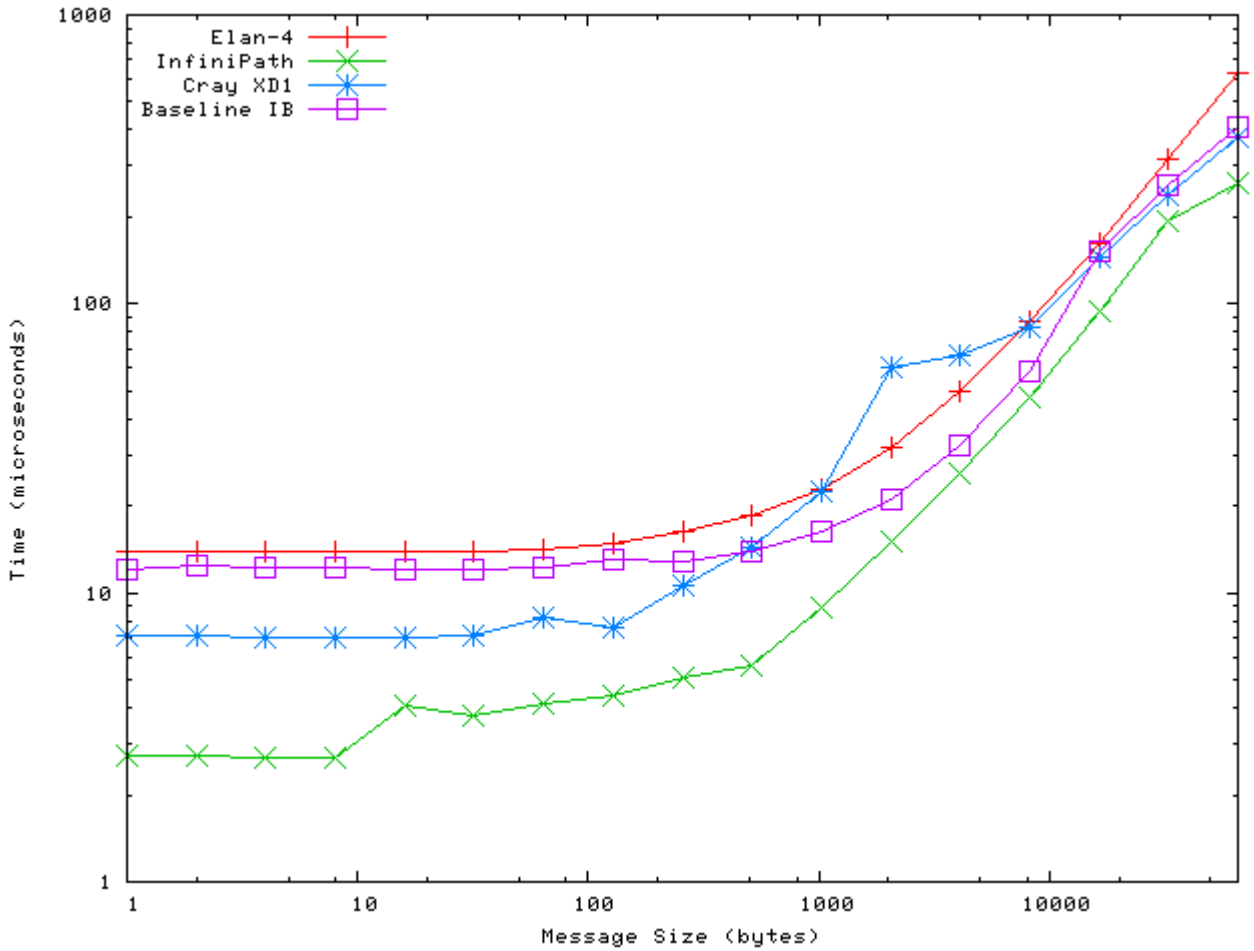
Sandia National Laboratories

32-Node Exchange Bandwidth



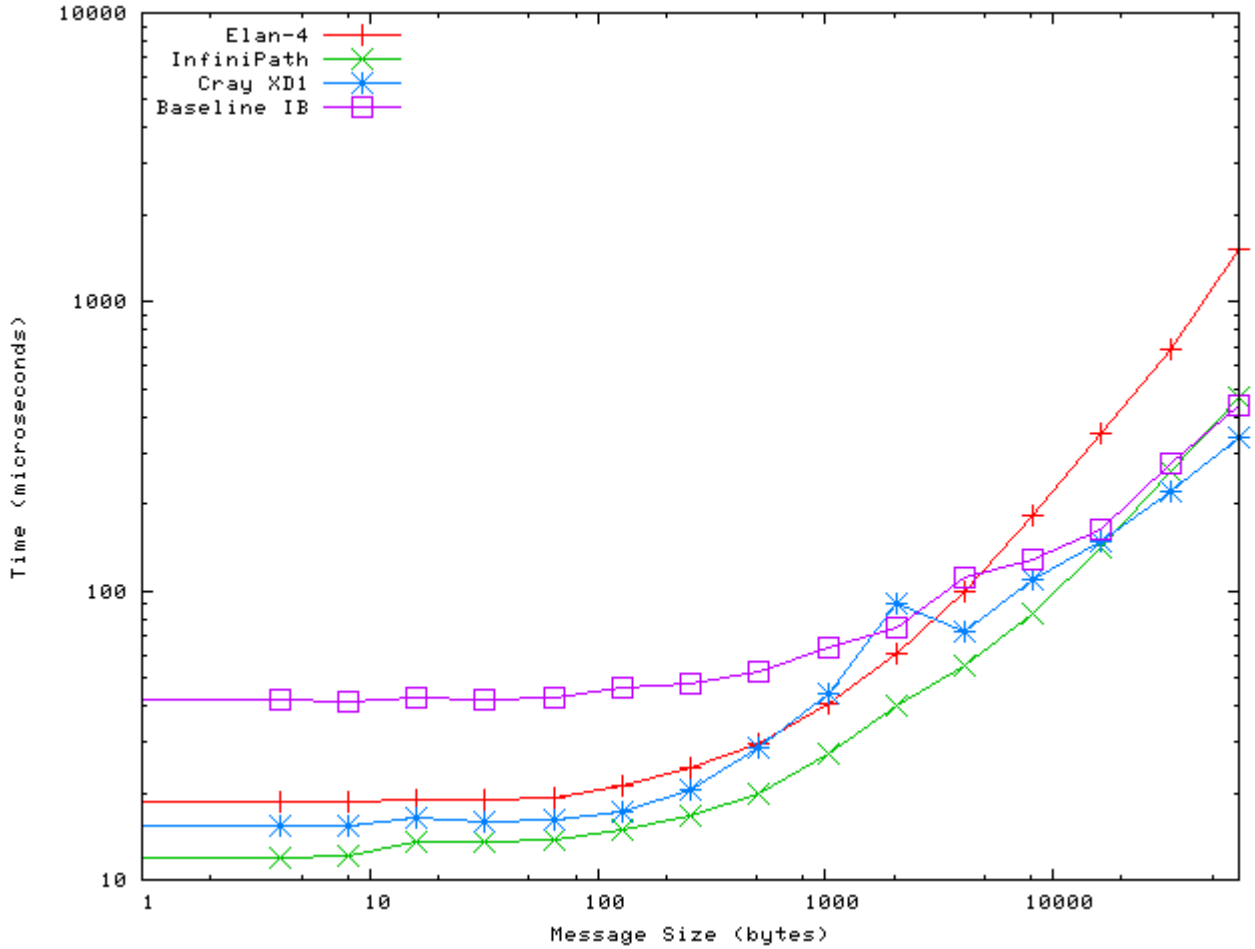
Sandia National Laboratories

32-Node Broadcast

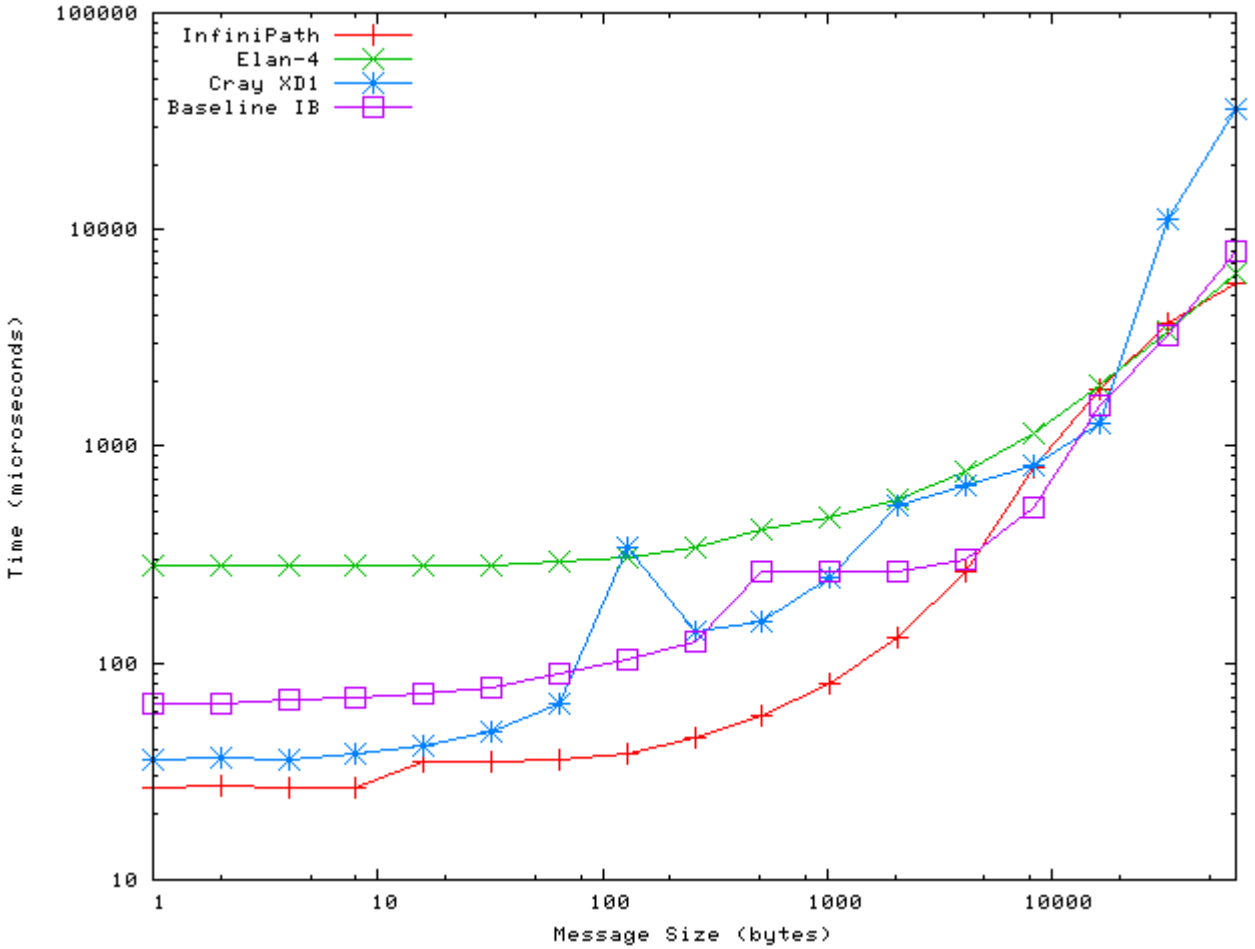


Sandia National Laboratories

32-Node Allreduce

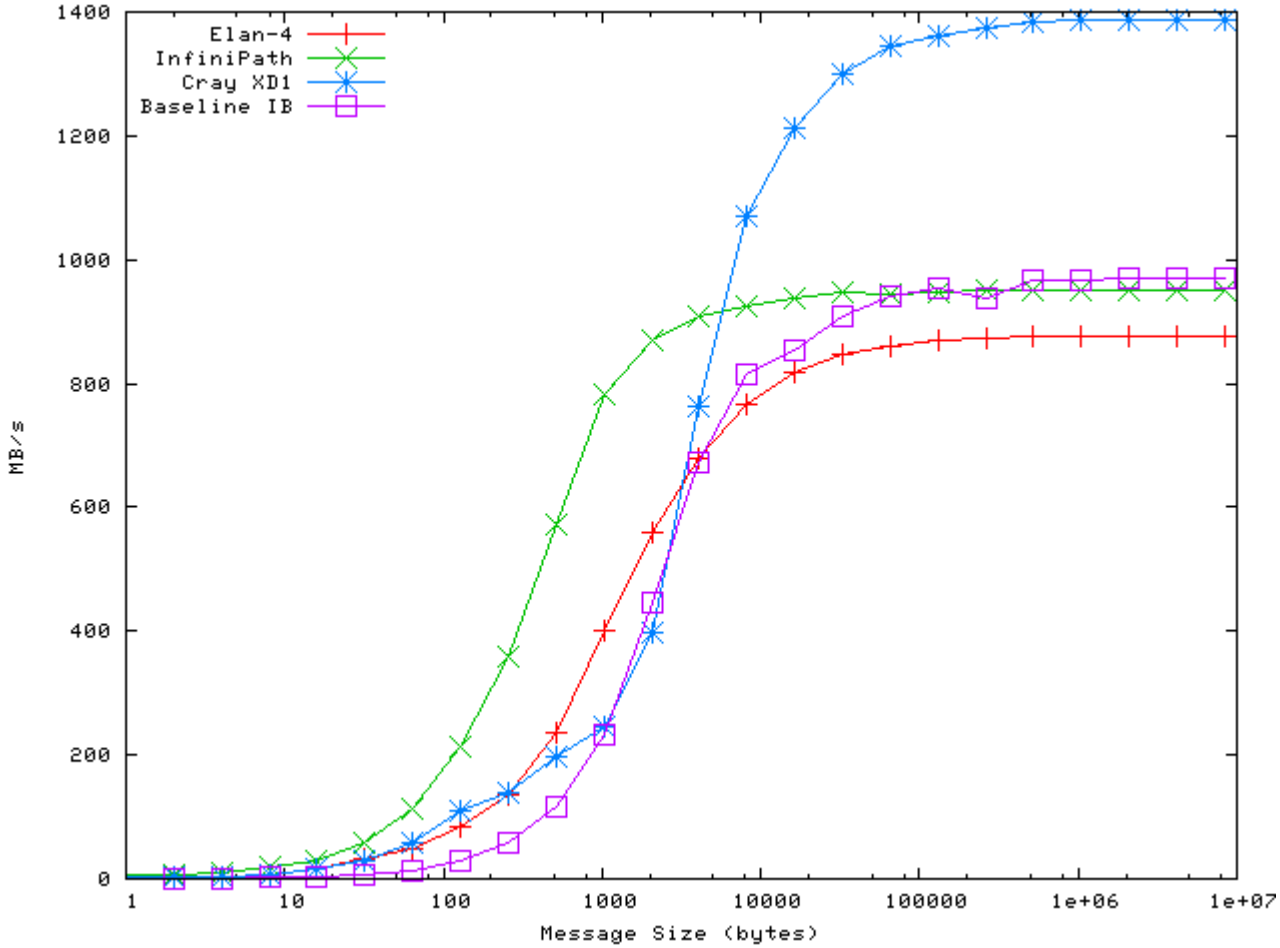


32-Node Alltoall



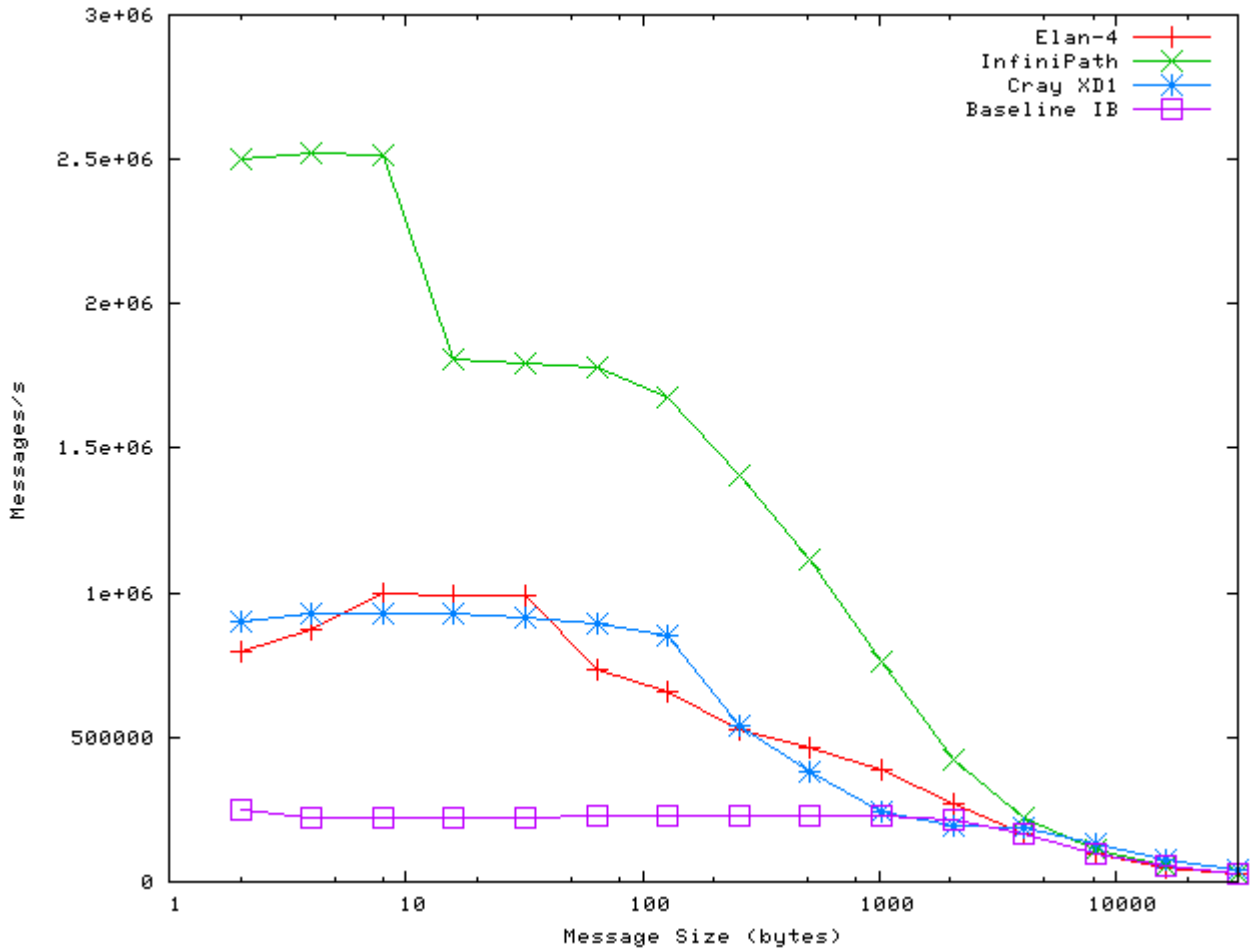
Sandia National Laboratories

Streaming Bandwidth



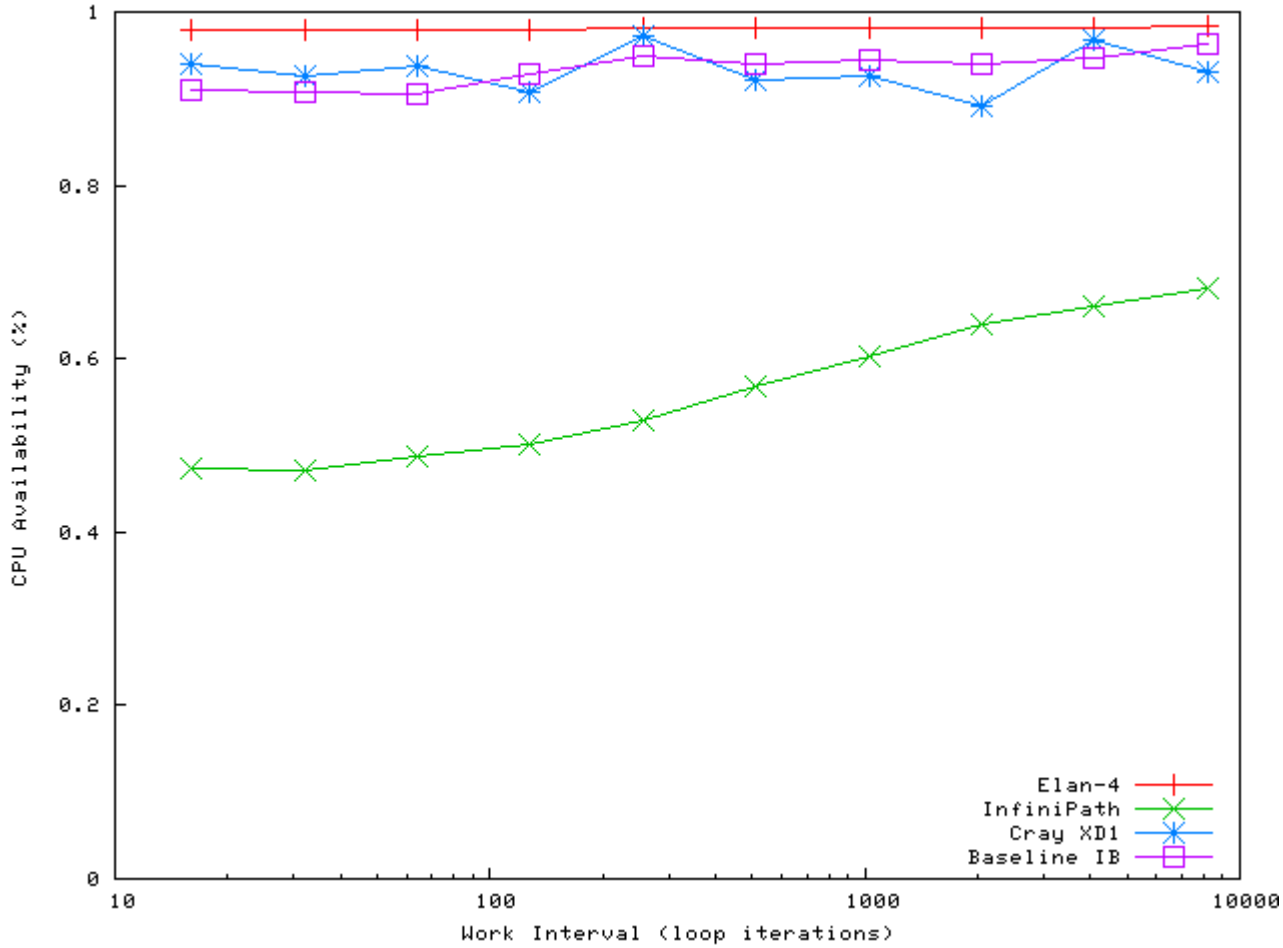
Sandia National Laboratories

Message Rate



Sandia National Laboratories

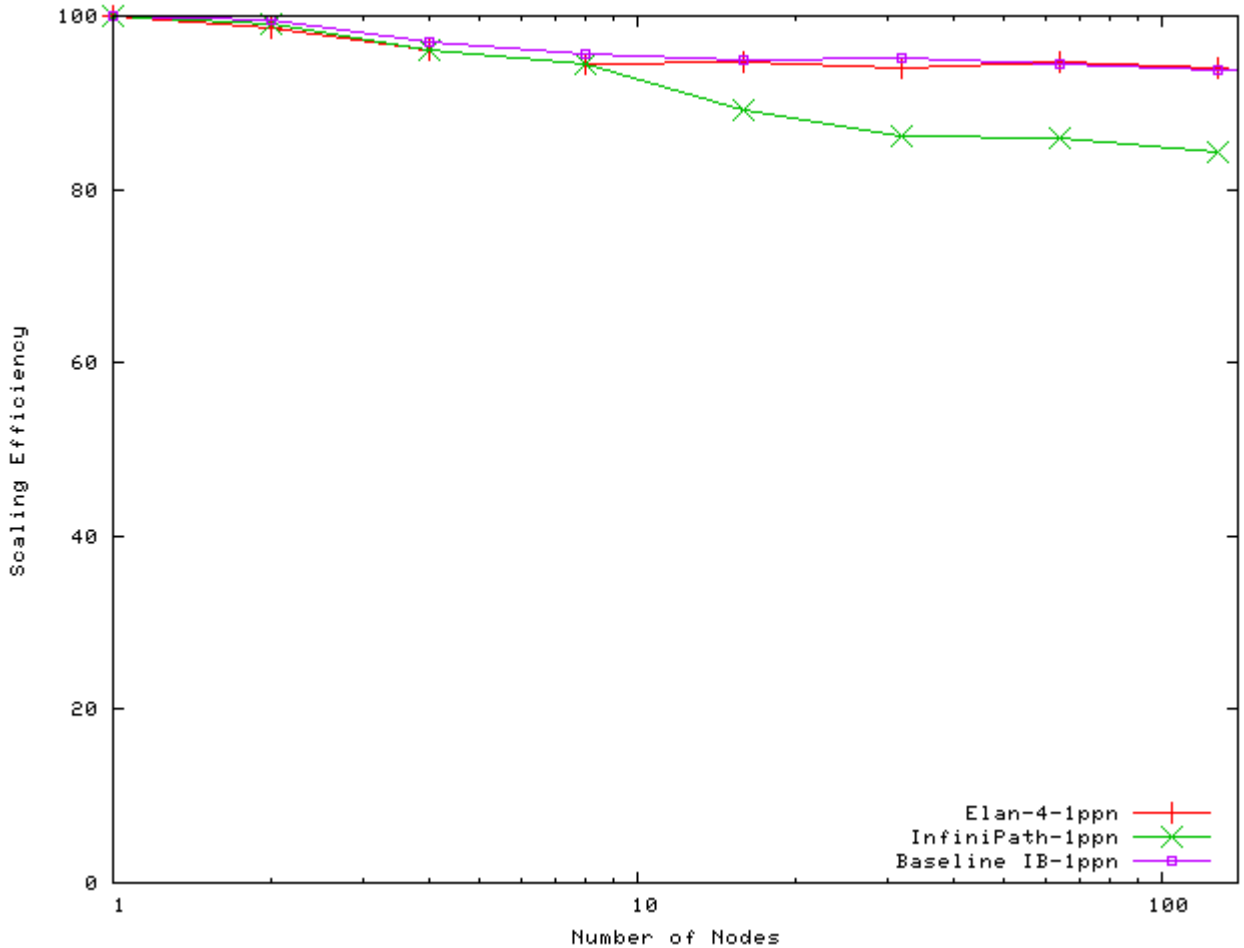
CPU Availability



Sandia National Laboratories

LAMMPS-2001 Efficiency

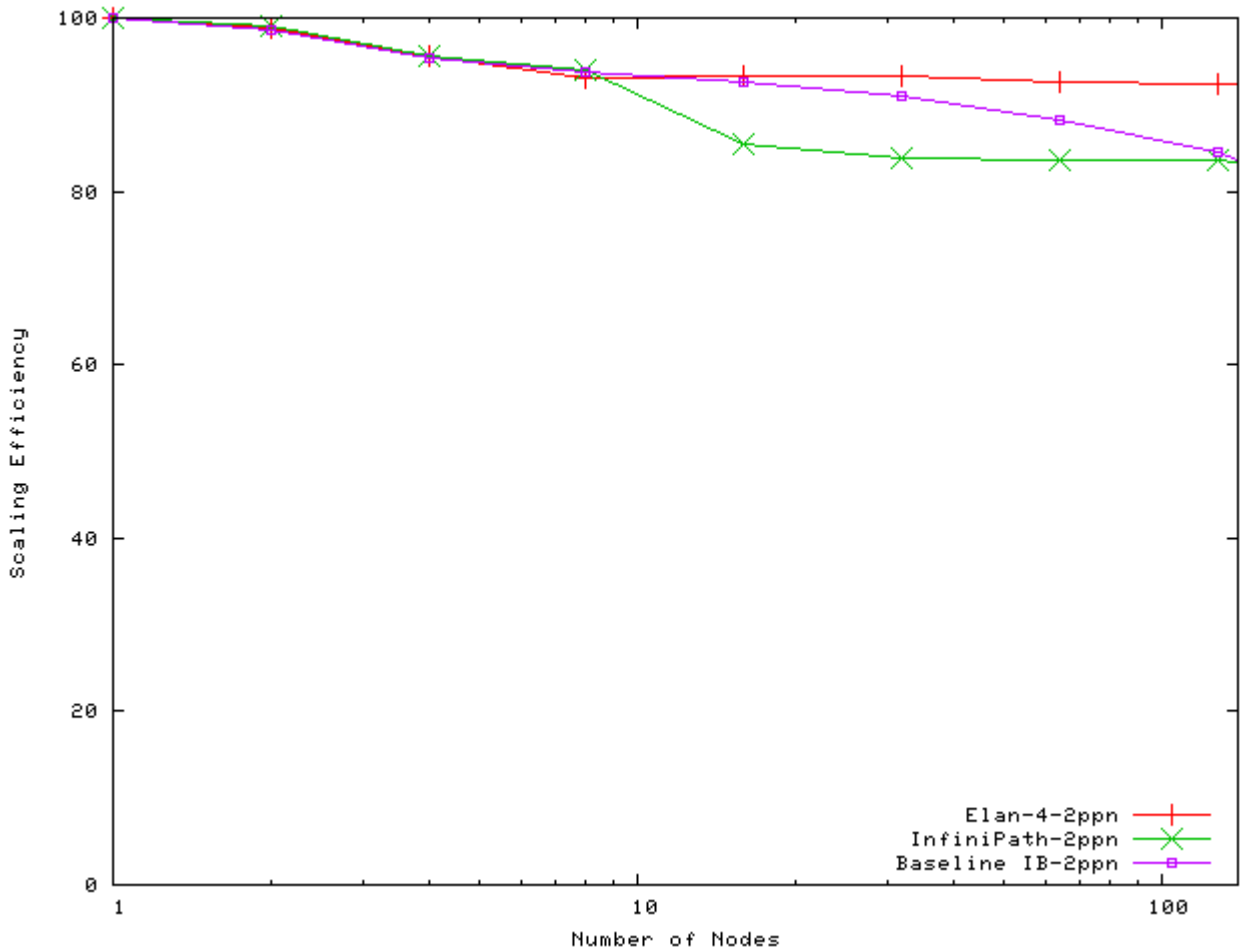
(1 process per node)



Sandia National Laboratories

LAMMPS-2001 Efficiency

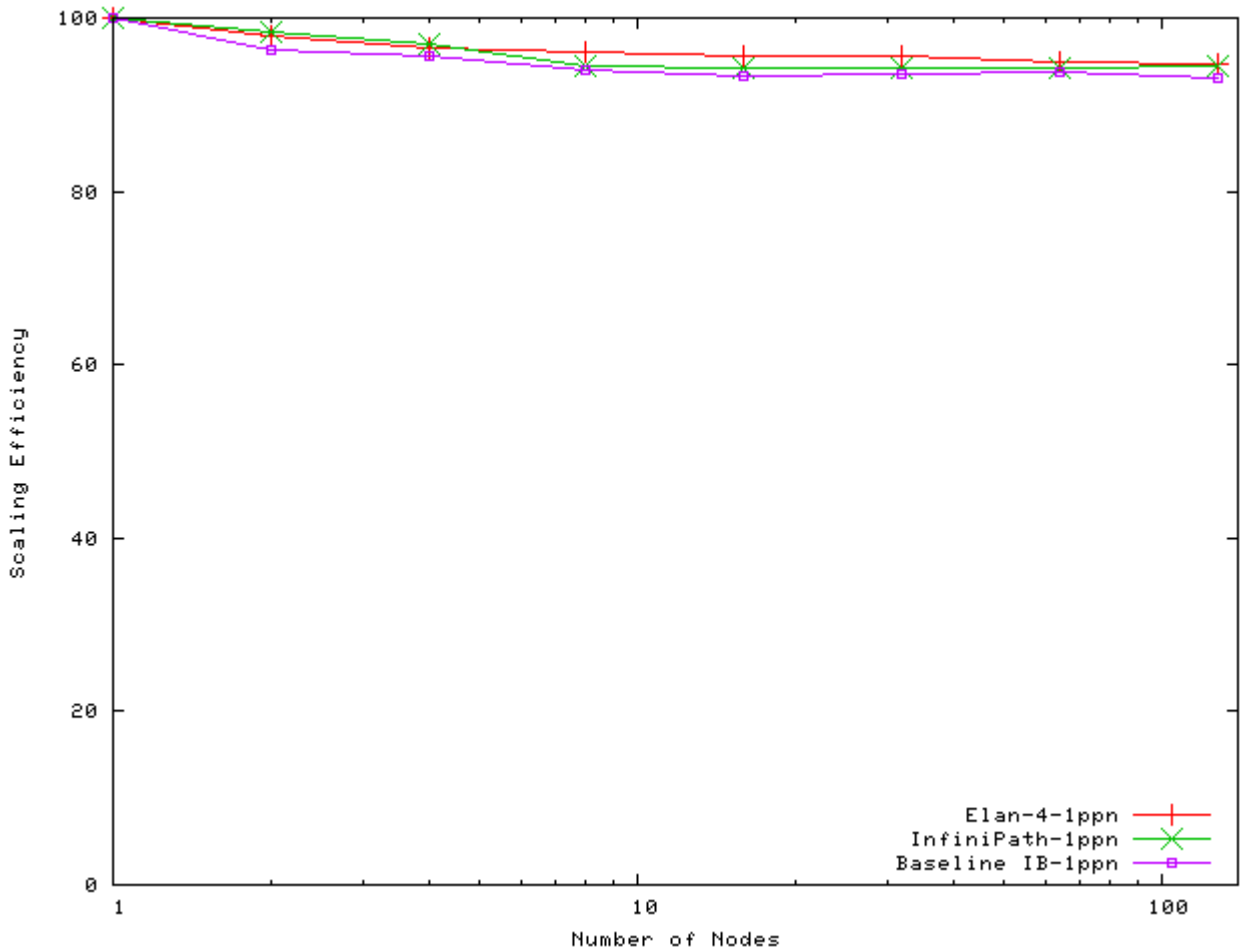
(2 processes per node)



Sandia National Laboratories

LAMMPS-2005 Efficiency

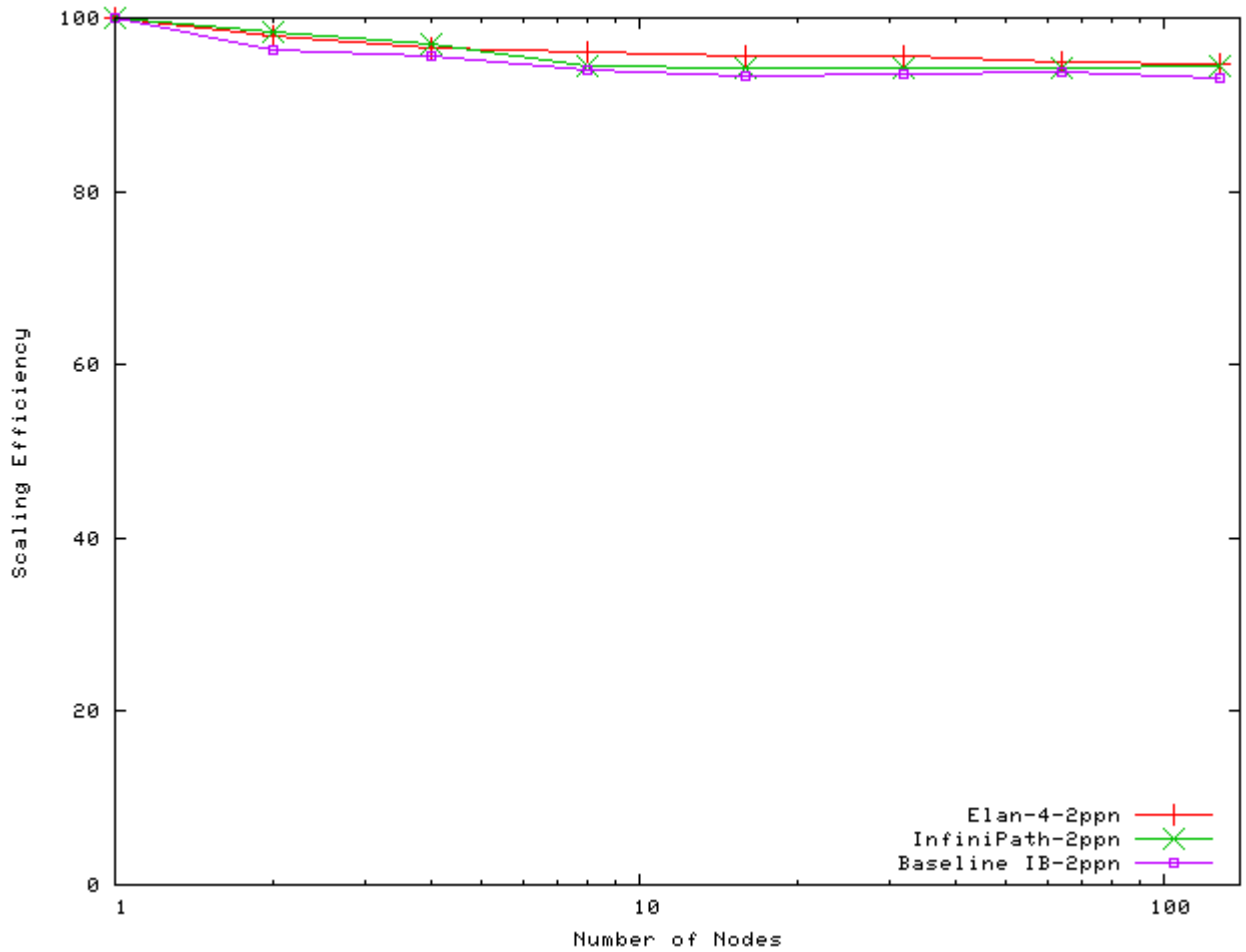
(1 process per node)



Sandia National Laboratories

LAMMPS-2005 Efficiency

(2 processes per node)



Sandia National Laboratories

Conclusions

- InfiniPath and RAI demonstrate good performance relative to other established technologies
- Both demonstrate better latency performance than pure commodity IB NICs
- Message rate is also significantly better
- Traditional micro-benchmarks do not expose the drawbacks of using host CPU(s) for network functionality

Sandia National Laboratories

Future Work

- More sophisticated micro-benchmarks
  - Message rate
  - Impact of CPU availability
  - MPI queue traversal
- Real application analysis

Sandia National Laboratories

Acknowledgments

- Greg Lindahl
  - For telling us how the InfiniPath NIC didn't work
- AMD Developer Center (http://devcenter.amd.com)
  - For access to the Emerald platform
- ORNL National Center for Computational Sciences
  - For access to their Cray XD1

Sandia National Laboratories